DSC680 - Project 2

David Suffolk

Overview

- Introduction of Problem
- Data Source
- Data Exploration
- Machine Learning Models
- Conclusion

Introduction of Problem

- Diseases spread from animals to humans (Zoonotic Diseases)
- Important to understand animal diseases
- Data Science can help to find answers to possible questions
    - What factors can help predict presence of blood pathogens?
    - Are blood samples required?
    - Can a location of an outbreak be identified?

Data Source

- Data collected on Alaskan crows from 2007-2008
- Alaskan Science Center
- Concern of growth in beak deformities
    - Avian Keratin Disorder (AKD)

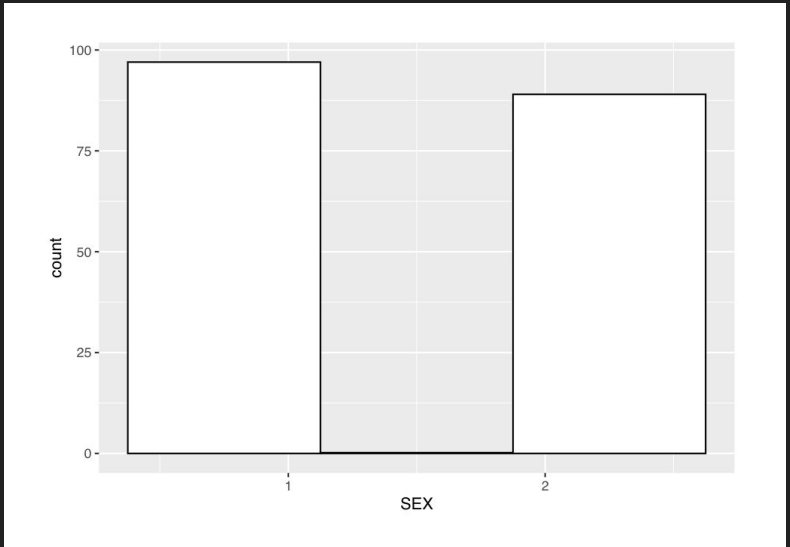
Data Source

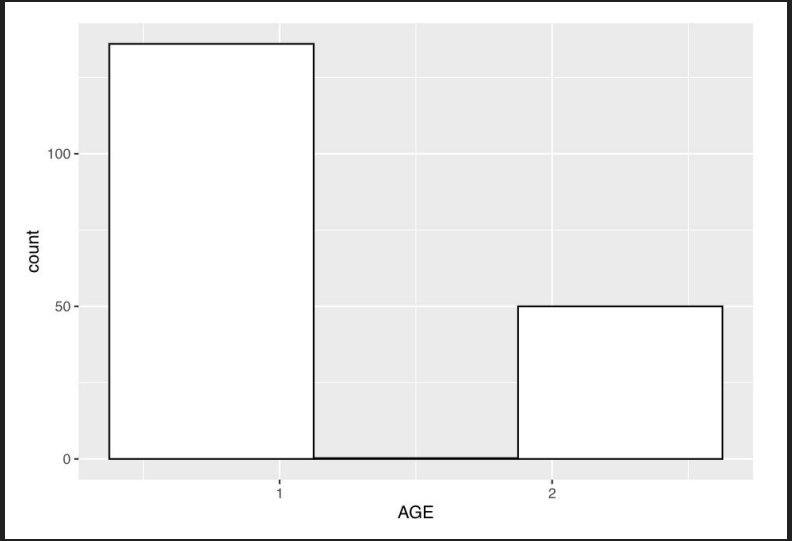
- Samples of over 180 crows
- Biological data and blood samples
- Studied for presence of three blood pathogens
  - Leucocytozoon parasite infection (LEUC)
  - Haemoproteus parasite infection (HAEM)
  - Plasmodium parasite infection (PLAS)

Data Exploration

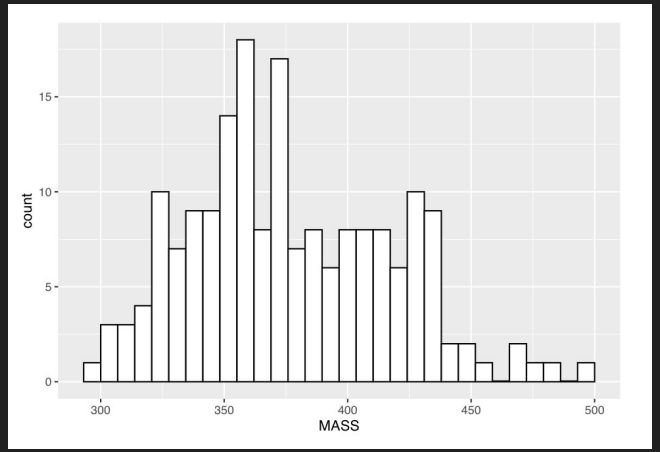
- All functions, graphs, and visuals created in R
- Six locations
  - One-hot encoding used for the machine learning algorithms
- Two tests for each pathogen
  - First test result was used for the majority of analysis

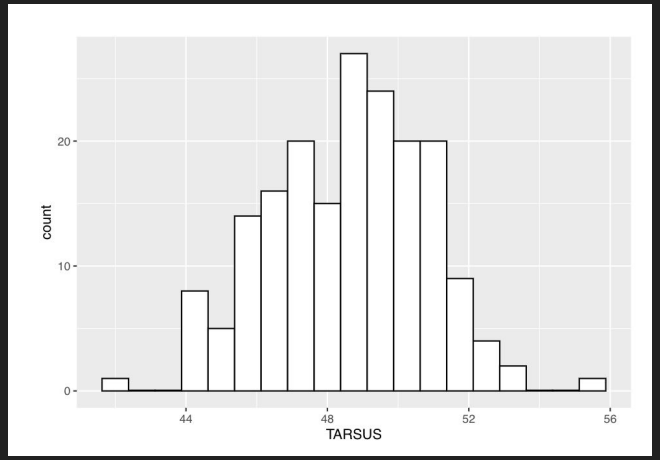
Data Exploration



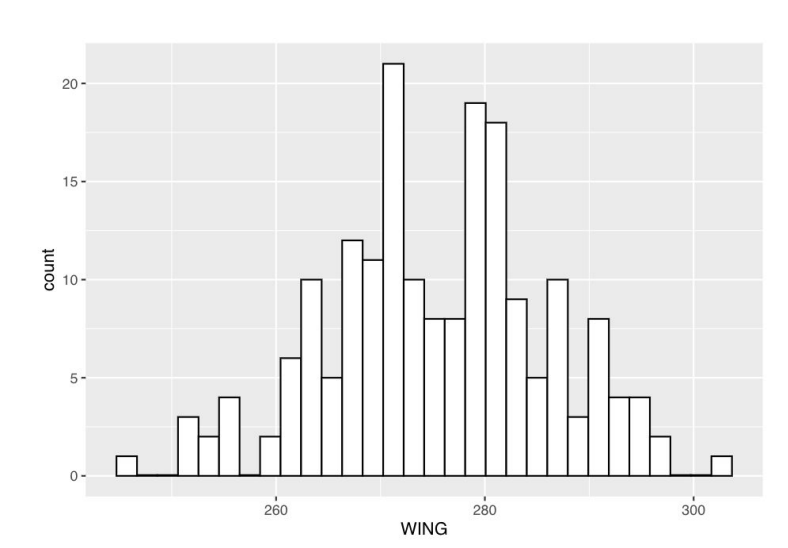


Data Exploration

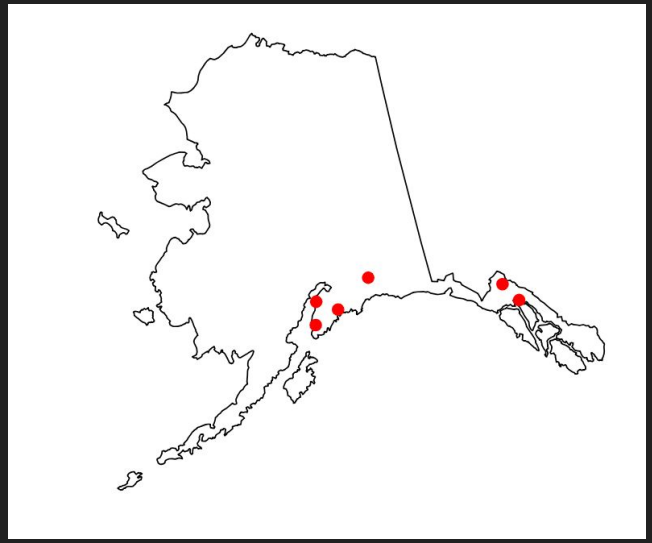


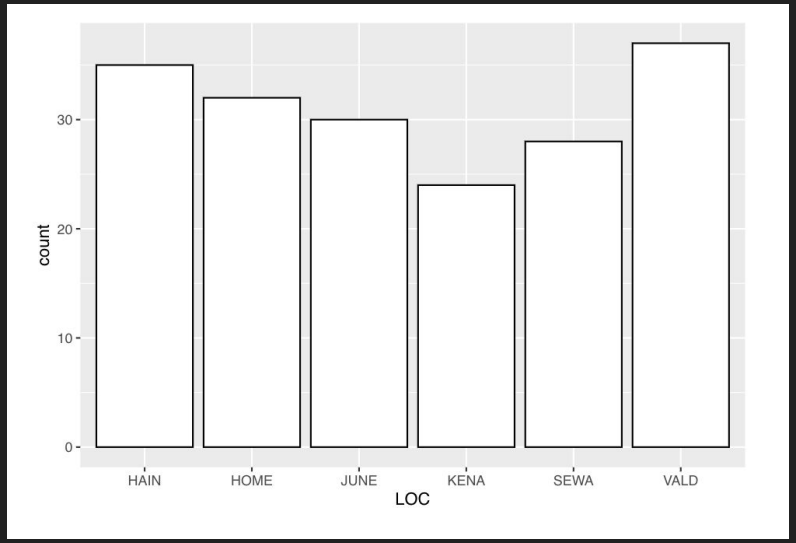


Data Exploration

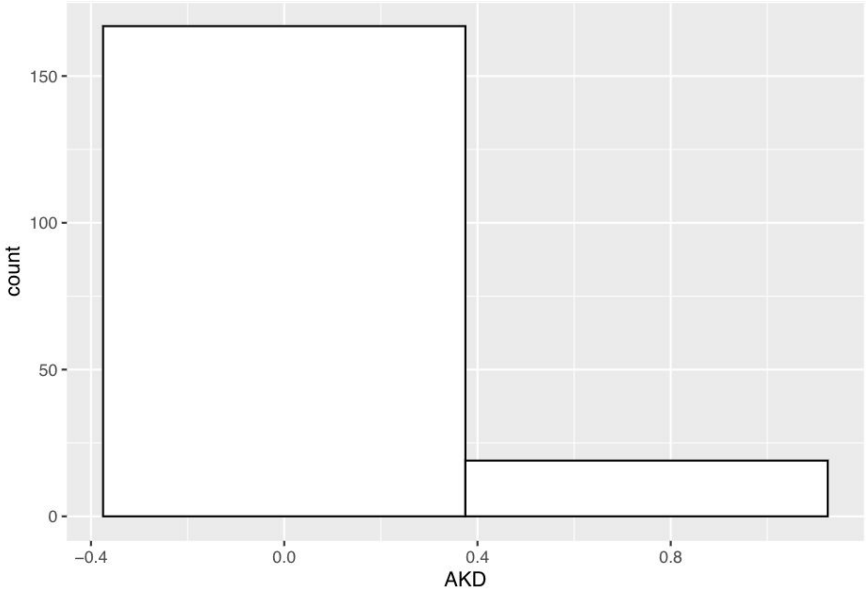


Data Exploration

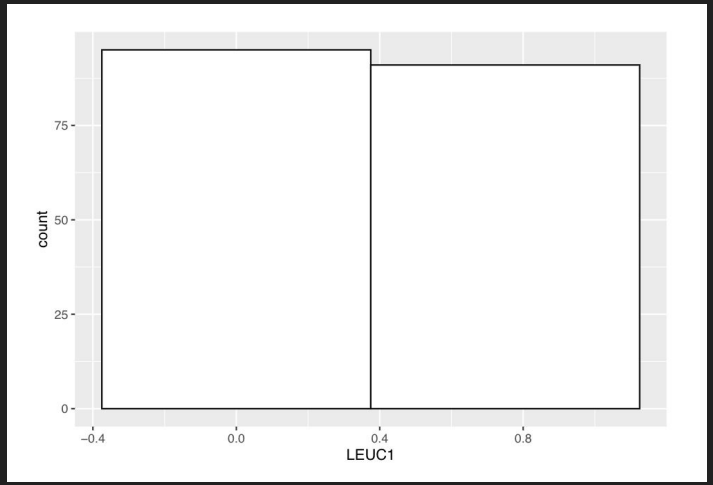


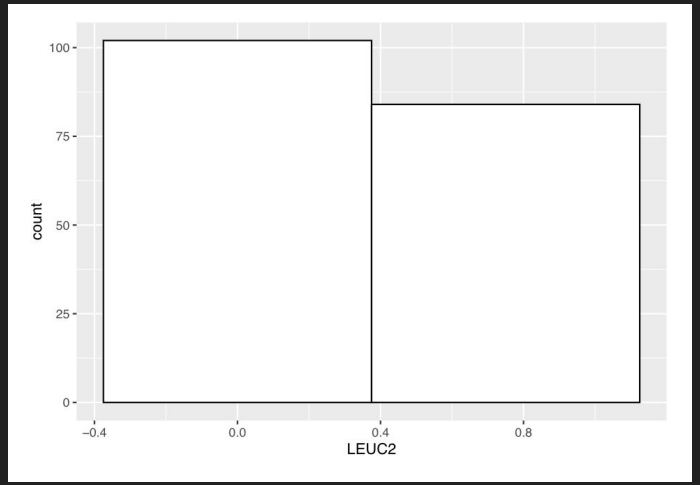


Data Exploration

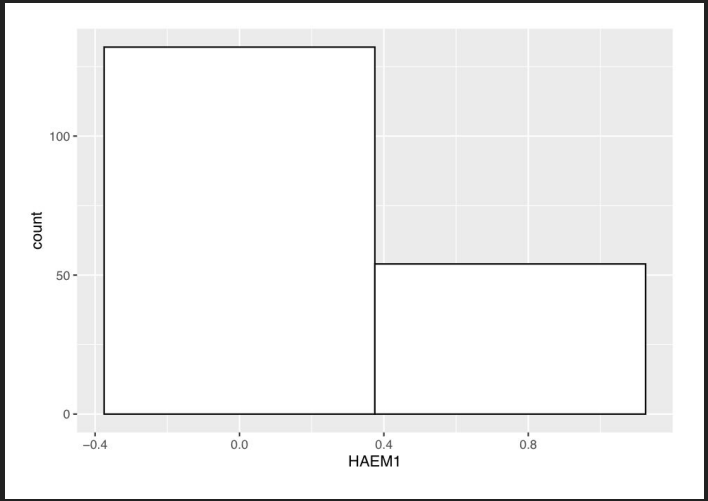


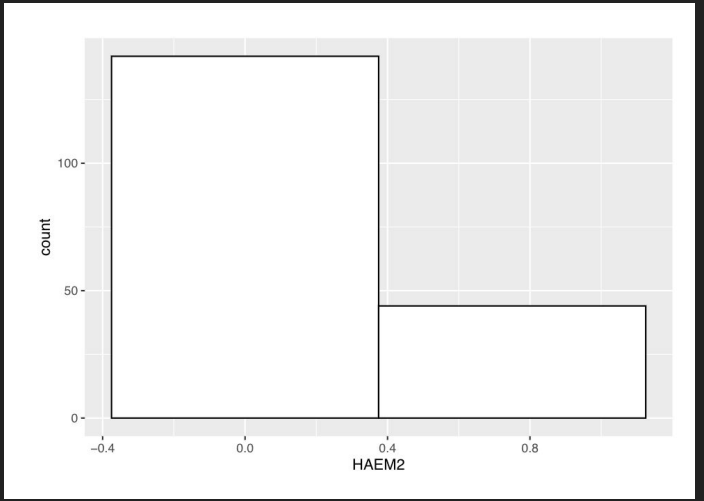
Data Exploration



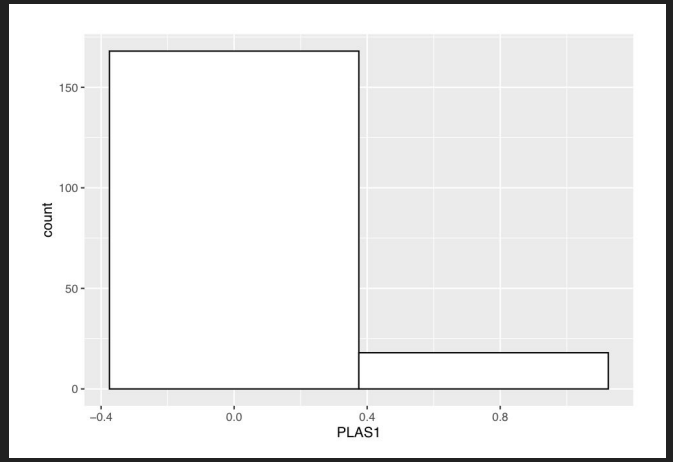


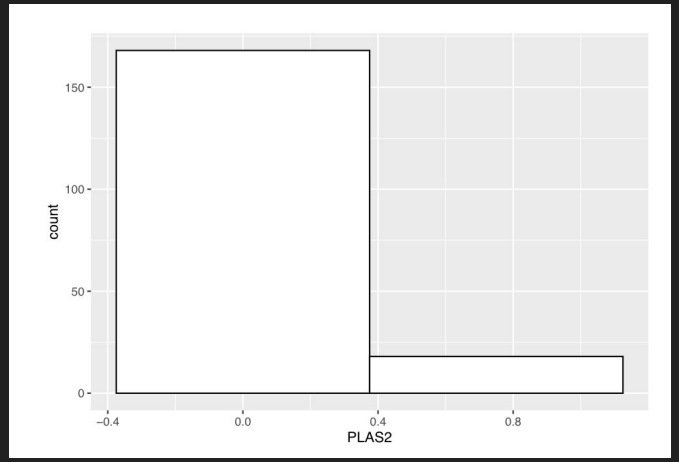
Data Exploration



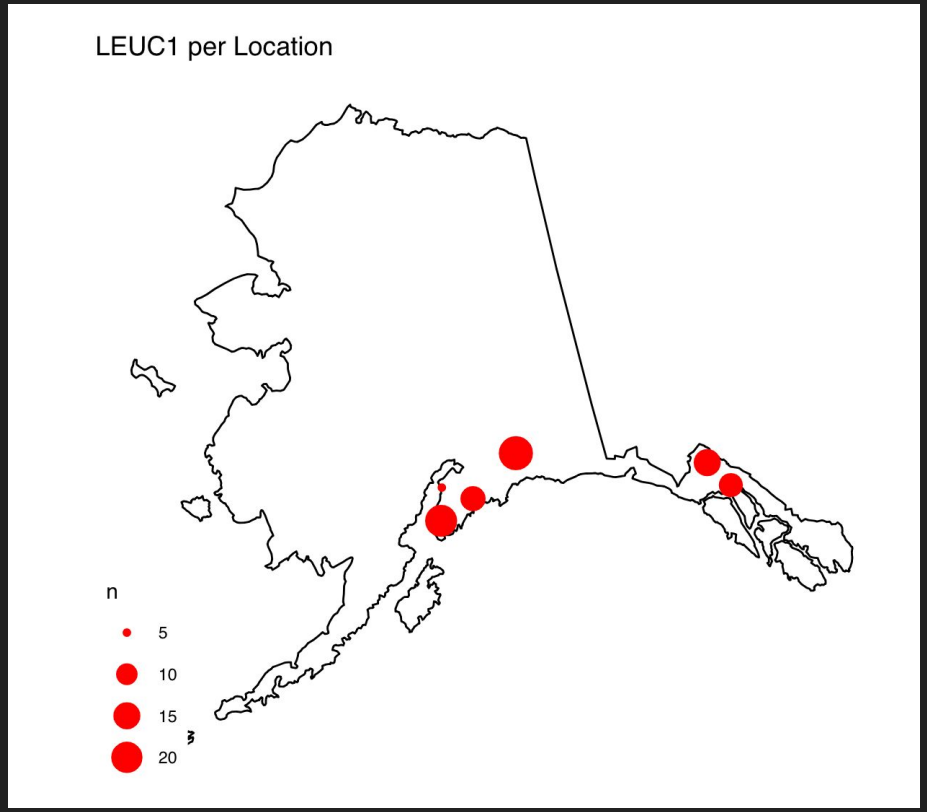


Data Exploration

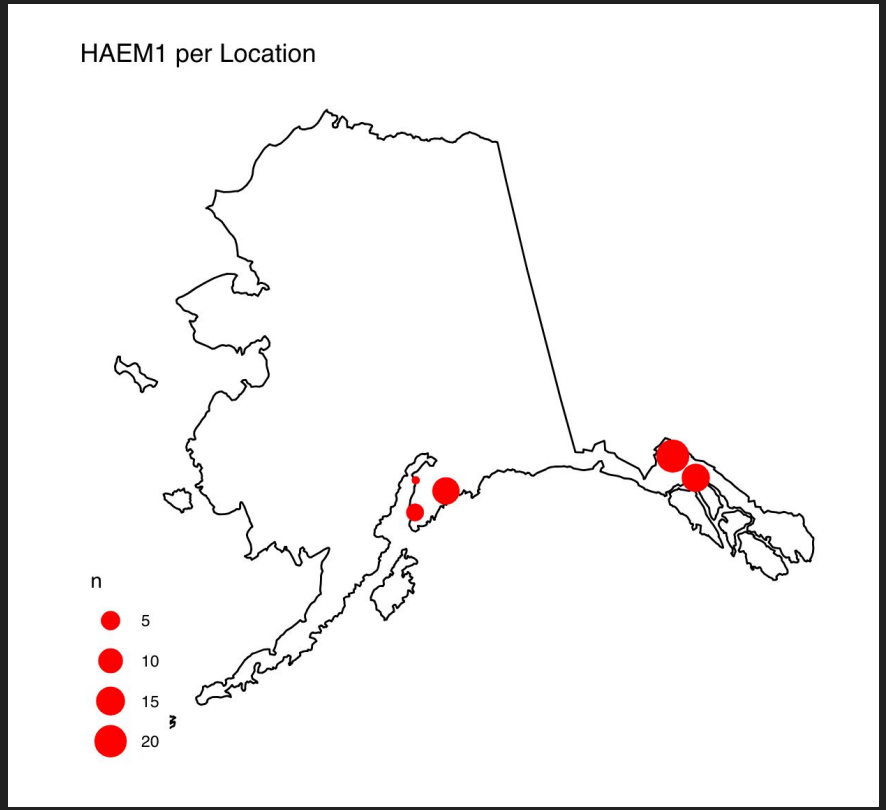




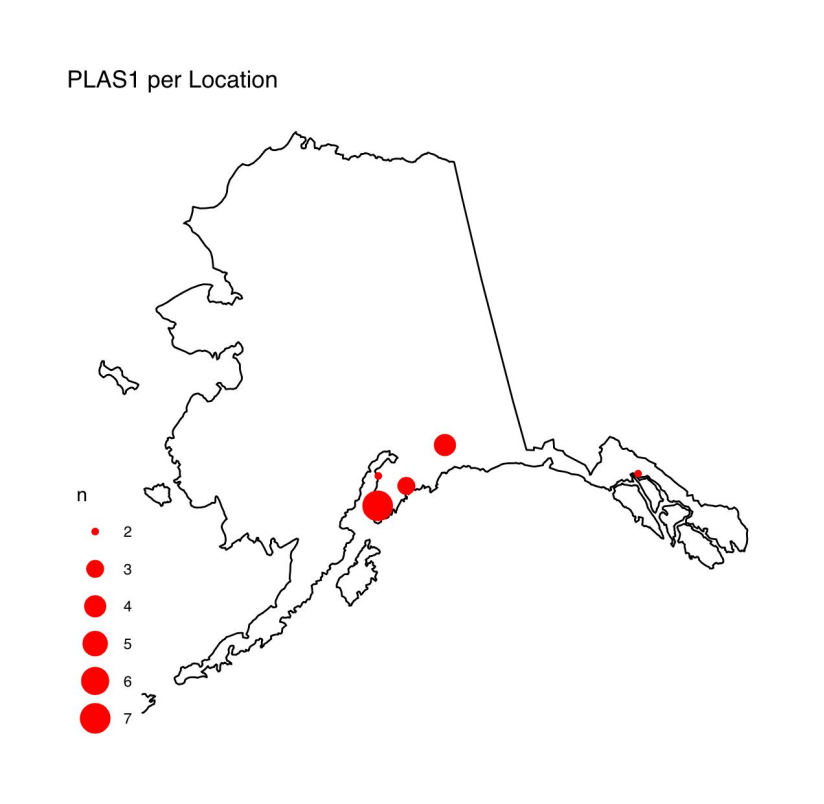
Data Exploration



Data Exploration



Data Exploration



Machine Learning Models

- Target Variables
  - AKD
  - LEUC1
  - HAEM1
  - PLAS1
- No significant covariance or correlation between target variables and any feature variables
- Linear Regression Model found no significant relationships

Machine Learning Models

- 3 Algorithms Used
  - Logistic Regression
  - Decision Tree
  - Naive Bayes
- Target variables are binary

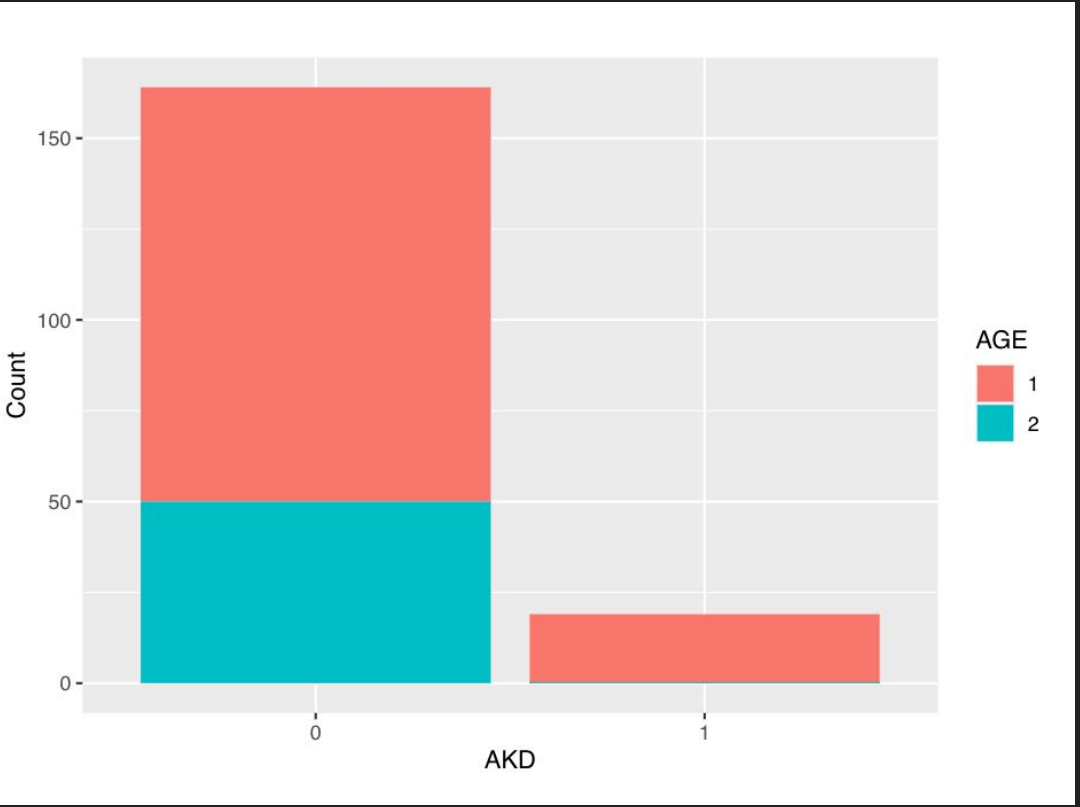
Logistic Regression - AKD

```
##                    2.5 %       97.5 %
## (Intercept) -2634.1900120 2.703235e+03
## SEX            -1.7216634 6.205950e-01
## AGE         -2687.5585861 2.649684e+03
## TARSUS         -0.5595384 1.466764e-01
## WING           -0.1411362 3.639016e-02
## MASS           -0.0105768 4.295700e-02
## LEUC1          -1.4942911 9.954247e-01
## HAEM1          -0.4354375 2.801334e+00
## PLAS1          -1.1808293 2.272975e+00
## LOC.SEWA        0.3159412 5.206111e+00
## LOC.KENA        0.4843588 5.526163e+00
## LOC.VALD       -0.9485564 3.915921e+00
## LOC.HAIN       -3.1540757 2.962598e+00
## LOC.JUNE       -1.3598649 3.762584e+00
## LOC.HOME               NA          NA
```

	FALSE	TRUE
0	121	2
1	12	2

Accuracy = 89.78%

Logistic Regression - AKD



Logistic Regression - LEUC1

```
##                    2.5 %       97.5 %
## (Intercept) -17.71263917  5.909124006
## SEX          -1.31020091  0.040505299
## AGE          -0.64902989  0.988229952
## TARSUS       -0.16748996  0.251269057
## WING         -0.01898793  0.079701079
## MASS         -0.02030441  0.004506129
## AKD          -1.39307227  0.850198909
## HAEM1        -0.25512635  1.438850338
## PLAS1        -1.41951384  0.862883216
## LOC.SEWA     -2.33247532  0.224501978
## LOC.KENA     -3.49483430 -0.744937835
## LOC.VALD     -1.24105373  0.962057262
## LOC.HAIN     -2.60165698 -0.162969019
## LOC.JUNE     -2.73548874 -0.310925938
## LOC.HOME              NA           NA
```

	FALSE	TRUE
0	49	21
1	21	47

Accuracy = 69.57%

Logistic Regression - HAEM1

```
##                    2.5 %       97.5 %
## (Intercept) -2.555382e+01 6.897946e+00
## SEX         -9.989220e-01 7.764953e-01
## AGE          2.875900e-01 2.583641e+00
## TARSUS      -1.322820e-01 4.189151e-01
## WING        -5.266542e-02 7.420183e-02
## MASS        -2.597683e-02 5.088832e-03
## LEUC1       -2.017605e-01 1.558561e+00
## AKD         -8.411244e-01 2.107398e+00
## PLAS1       -4.370732e+03 4.334582e+03
## LOC.SEWA    -5.049756e-01 2.711085e+00
## LOC.KENA    -4.645821e+00 2.663481e-01
## LOC.VALD    -3.216350e+03 3.179455e+03
## LOC.HAIN     3.676849e-01 3.202834e+00
## LOC.JUNE    -1.759188e-01 2.749479e+00
## LOC.HOME              NA           NA
```

	FALSE	TRUE
0	90	8
1	12	28

Accuracy = 85.5%

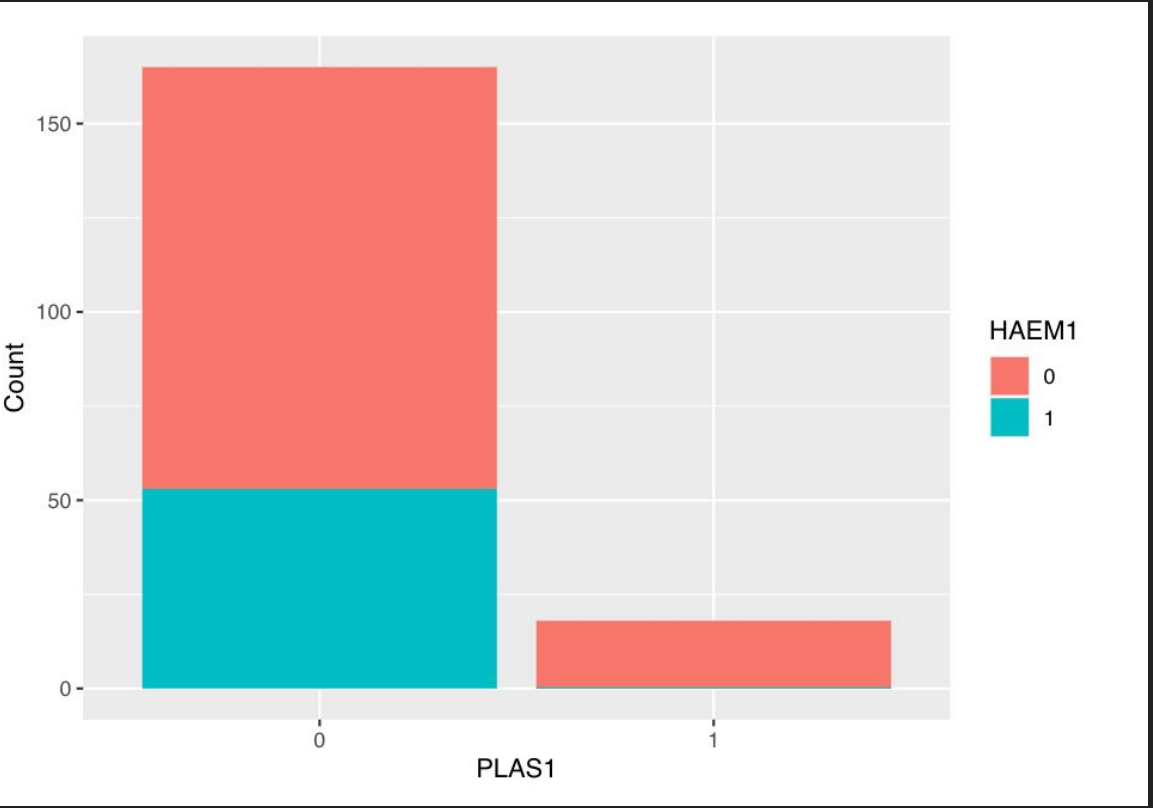
Logistic Regression - PLAS1

```
##                    2.5 %        97.5 %
## (Intercept) -4.024072e+01  3.899895e+00
## SEX         -1.182815e+00  1.132212e+00
## AGE         -1.785850e+00  1.738433e+00
## TARSUS      -5.059899e-01  2.209062e-01
## WING         2.125323e-02  2.226448e-01
## MASS        -5.219369e-02 -1.072603e-03
## LEUC1       -1.421506e+00  9.861846e-01
## HAEM1       -4.066102e+03  4.028546e+03
## AKD         -7.635654e-01  2.683554e+00
## LOC.SEWA    -1.478637e+00  2.524080e+00
## LOC.KENA    -3.550206e+00  5.054281e-01
## LOC.VALD    -2.356527e+00  5.392387e-01
## LOC.HAIN    -4.900745e+03  4.865228e+03
## LOC.JUNE    -2.369083e+00  1.309431e+00
## LOC.HOME              NA            NA
```

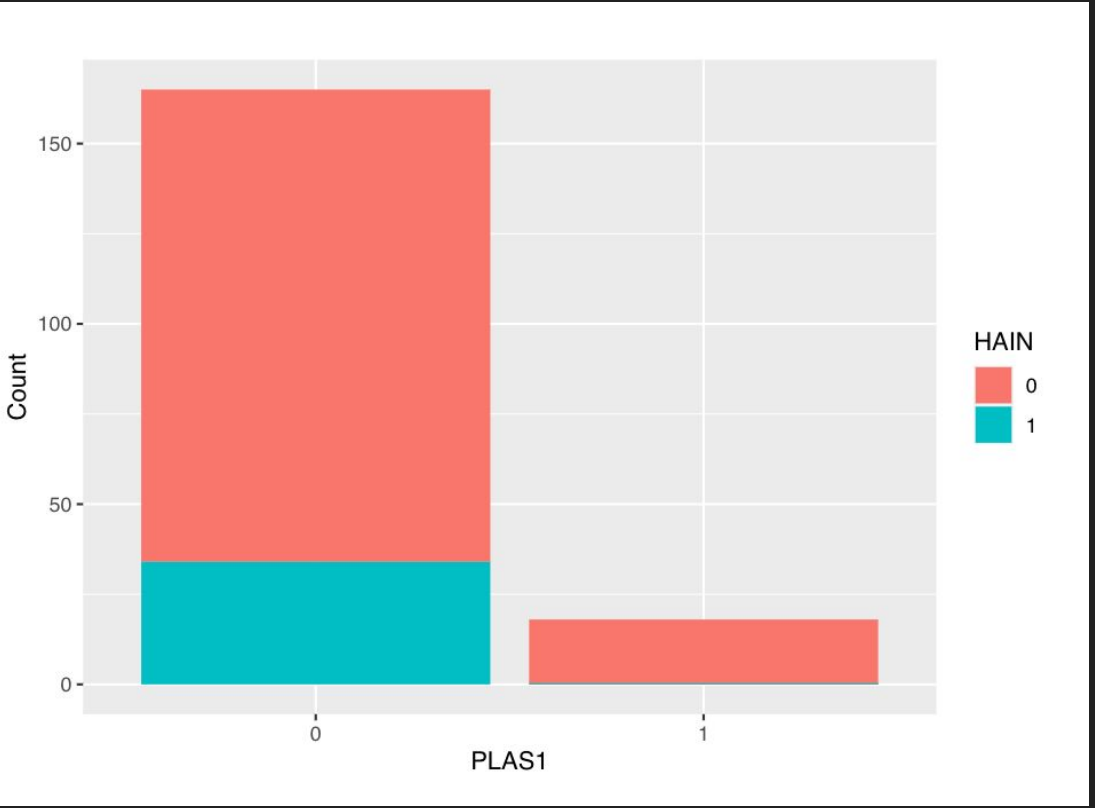
	FALSE	TRUE
0	123	1
1	10	4

Accuracy = 92%

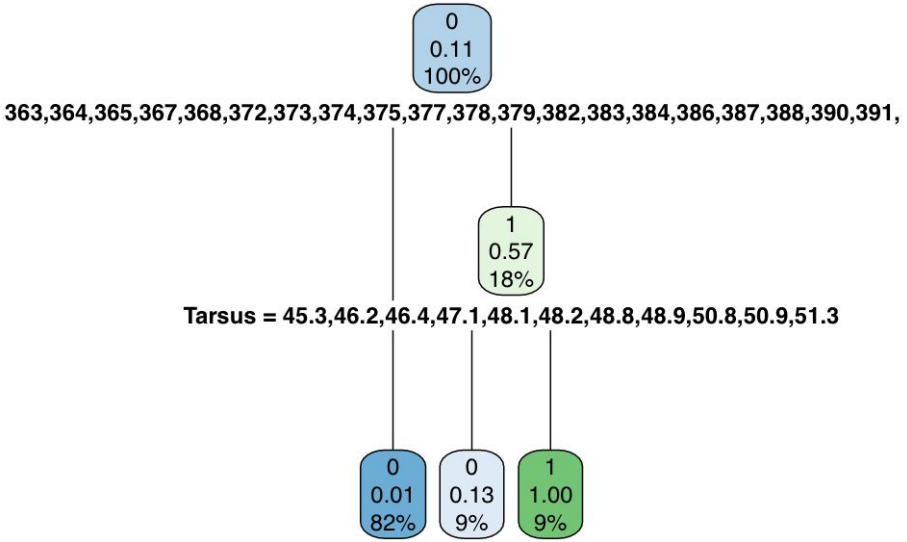
Logistic Regression - PLAS1



Logistic Regression - PLAS1



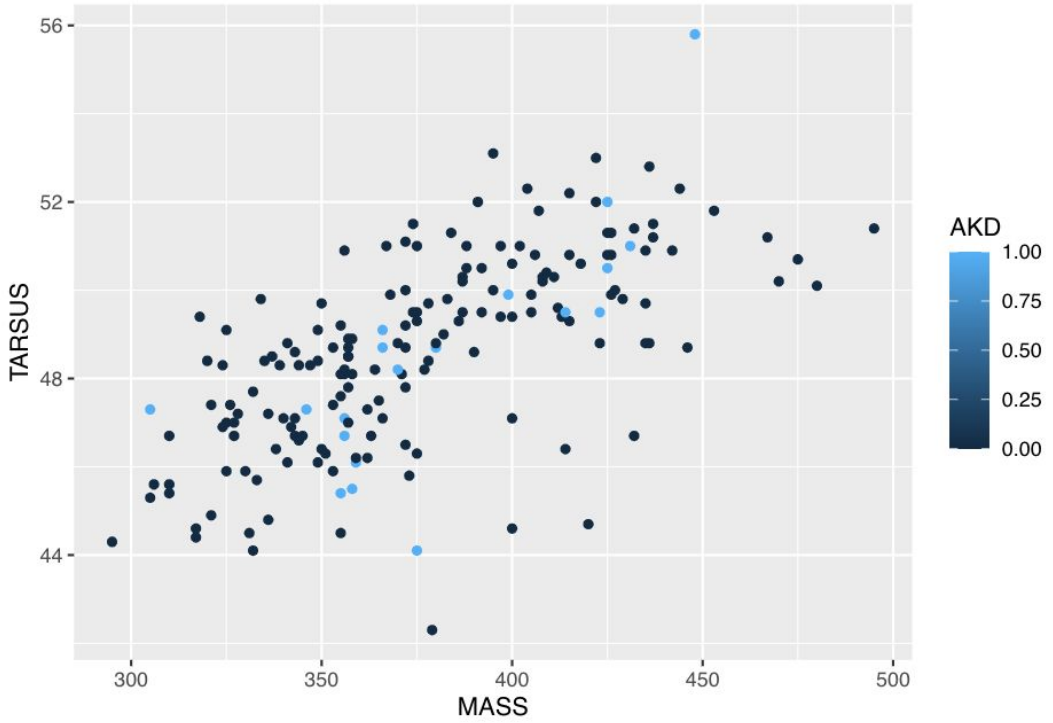
Decision Tree - AKD




```
p <- predict(dtm, data_test, type="class")
confMat <- table(data_test$AKD,p)
accuracy <- sum(diag(confMat))/sum(confMat)
return (accuracy*100)
```

```
## [1] 94.73684
```

Decision Tree - AKD



Decision Tree - Pathogens

- LEUC1 - 47.37%
- HAEM1 - 63.16%
- PLAS1 - 73.68%

Naive Bayes - AKD

```
##
## p2    0  1
##   0 28  4
##   1  0  0
```

```
## [1] 87.5
```

Naive Bayes - LEUC1

```
#
# p2  0 1
#   0 8 8
#   1 8 8
```

```
## [1] 50
```

Naive Bayes - HAEM1

```
p2   0  1
  0 16  4
  1  4  8
```

```
## [1] 75
```

Naive Bayes - PLAS1

p2	0	1
0	25	5
1	2	0

```
[1] 78.125
```

Machine Learning Models

	Target Variable			
Algorithm	**AKD**	**LEUC1**	**HAEM1**	**PLAS1**
Logistic Regression	89.78%	69.57%	85.50%	92%
Decision Tree	94.70%	47.37%	63.16%	73.68%
Naive Bayes	87.50%	50%	75%	78.13%

Conclusion

- AKD has potential to be predicted reliably
  - Age, mass, tarsus
- Logistic Regression held potentially strong results
  - AKD
  - PLAS1
  - HAEM1